Universal Source Polarization and Sparse Recovery

Emmanuel Abbe
LCM - EPFL
Email: emmanuel.abbe@epfl.ch

***Abstract*—Polar codes allow to perform lossless compression of i.i.d. sources at the lowest rate with low encoding and decoding complexity. In this paper, it is shown that for binary sources, there exist "universal polar codes" which can compress any source of low enough entropy, without requiring knowledge of the source distribution. While this result does not extend to q-ary sources, it is shown how it extends to q-ary sources which belong to a restricted family. An analogy between this family and BECs in channel polarization is discussed. Finally, an application of the universal source polarization results to sparse data recovery is proposed.**

I. Introduction

Arıkan shows in [1] that an arbitrary binary input discrete memoryless channel W can be polarized as follows: n independent uses of W can be transformed into n successive uses of synthesized binary input channels that have (except for a vanishing fraction) a uniform mutual information[1] which tends to 0 or 1 (with n). In [4], this result is extended to q-ary input alphabets where q is prime.

Theorem 1 ([4])**.** *Let W be a q-ary input discrete memoryless channel (q-DMC) with q prime, $n = 2^\ell$ and let $U^n = (U_1, \dots, U_n)$ be i.i.d. uniform random variables on $\mathbb{F}_q$. Let $X^n = U^n G_n$, where $G_n = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]^{\otimes \ell}$, and let Y^n be the output of n independent uses of W when the input is X^n. Then, for any $\delta \in (0,1)$,*

$$\frac{1}{n}|\{i : I(U_i; Y^n U^{i-1}) > 1 - \delta\}| \stackrel{n\to\infty}{\longrightarrow} I(W), \qquad (1)$$

where $I(W)$ is the uniform mutual information[2] of W.

Previous limit implies that for n large enough and except for a vanishing fraction of indices i, $I(U_i; Y^n U^{i-1})$ must be close to either 0 or 1. Hence, this suggests a coding scheme, where the information bits are sent uncoded on the indices i for which the channel is good, i.e., $I(U_i; Y^n U^{i-1})$ is close to 1, and where frozen bits that are predetermined are sent on the other channels. In [4], polar codes are proved to achieve[3] the uniform mutual information of any q-DMC, and the encoding and decoding complexity is shown to be $O(n \log n)$, which is remarkably low.

In [3], Arıkan uses an equivalent approach to perform lossless compression of memoryless sources.

[1]The mutual information of the channel evaluated with the uniform input distribution.

[2]Computations are made with the logarithm in base q.

[3]To show achievability, the speed of convergence to the polarized channels matters, and it is shown to be roughly $2^{-\sqrt{n}}$.

Theorem 2 ([3])**.** *Let X^n be n i.i.d. random variables with distribution p on $\mathbb{F}_q$ and let $U^n = X^n G_n$, where G_n is as defined in Theorem 1. Then, for any $\delta \in (0,1)$,*

$$\frac{1}{n}|\{i : H(U_i|U^{i-1}) > 1 - \delta\}| \stackrel{n\to\infty}{\longrightarrow} H(p), \qquad (2)$$

where $H(p)$ is the entropy of the distribution p.

Previous limit implies that for n large enough and except for a vanishing fraction of indices i, $H(U_i|U^{i-1})$ is close to either 0 or 1. Hence, the transformation G_n extracts the randomness in X^n, which is a priori uniformly distributed, into specific components indexed by i such that $H(U_i|U^{i-1})$ is close to 1, and puts all the randomness in these components. Lossless compression at lowest rate, $H(p)$, is then shown to be achievable with polar codes, and the complexity of encoding and decoding is again $O(n \log n)$.

In this work, we are interested in source polarization. In particular, we are interested in how sensitive are source polar codes to the knowledge of the source distribution. In applications, the source distribution is never perfectly known, and it is crucial that any used compression scheme does not rely too strongly on the knowledge of the source distribution. We will therefore consider the problem of constructing polar codes which can compress losslessly sources without requiring perfect knowledge of their distributions. We will then consider the problem of sparse data recovery, using polar codes. From the description of the source polarization theorem above, it makes sense that the signal acquisition problem is related to the compression problem: if we sense the signal U^n only in the components i for which $H(U_i|U^{i-1})$ is close to 1, we obtain a sampling of the signal which allows perfect recovery of the full signal, with a significantly reduced number of measurements. A first difference one can note between a compressed sensing setting [5], [6] and the polar code setting, is that in the latter setting, the signal is random. Hence, a natural question is to ask how sparsity, i.e., the property of having many components that are 0, is modeled for random signals, and how much the choice of a specific sparse probability distribution matters. Using our results on universal source polar codes, we then investigate how our approach compares to the approach of [5], [6], in terms of the number of measurements that it requires to allow perfect reconstruction.

II. Duality source-channel polarization

In this section, we clarify the relationship between Theorem 1 and Theorem 2. Let p be a distribution on $\mathbb{F}_q$ and consider using Theorem 1 for an additive noise channel, i.e., $Y = X \oplus Z$

for Z distributed under p and independent of X. We then have $Y^n = G_n U^n \oplus Z^n$ and

$$\begin{aligned} I(U_i; Y^n U^{i-1}) &= 1 - H(U_i|Y^n U^{i-1}) \\ &= 1 - H((G_n Y^n \ominus G_n Z^n)_i | Y^n (G_n Y^n \ominus G_n Z^n)^{i-1}) \\ &= 1 - H((G_n Z^n)_i | (G_n Z^n)^{i-1}). \end{aligned} \tag{3}$$

Equality (3) uses the fact that Y^n is independent of Z^n because U^n, and hence $G_n U^n$, are uniformly distributed over $\mathbb{F}_q$. We also use the fact that $G_n^{-1} = G_n$. Hence, Theorem 1 and (3) imply Theorem 2.

III. Mathematical preliminaries on ordering

Definition 1 (Measures)**.** Let a be a prime integer, $\mathbb{F}_a := \{0, 1, \dots, a-1\}$ and $\mathrm{M}(a)$ be the set of probability measures on $\mathbb{F}_a$. For any $k \in \mathbb{F}_a$, let

$$\widehat{\mathrm{M}}_k(a) := \{p \in \mathrm{M}(a) : p(i) = p(j),\, \forall i, j \neq k, p(k) \geq \frac{a-1}{a}\}$$

and $\widehat{\mathrm{M}}(a) := \cup_{k \in \mathbb{F}_a} \widehat{\mathrm{M}}_k(a)$. We refer to the measure in $\widehat{\mathrm{M}}(a)$ as the the spike measures.

Definition 2 (Matrices)**.** We denote by $\mathrm{Doub}(a)$ the set of doubly stochastic matrices of size $a \times a$, and by $\mathrm{Circ}(a)$ the set of circulant stochastic matrices of size $a \times a$.

Definition 3 (Orders)**.** We define

$$p_1 \prec_h p_2 \quad \text{if} \quad h(p_1) \geq h(p_2), \tag{4}$$

$$p_1 \prec_d p_2 \quad \text{if} \quad p_1 = D p_2 \text{ for } D \in \mathrm{Doub}(a), \tag{5}$$

$$p_1 \prec_c p_2 \quad \text{if} \quad p_1 = C p_2 \text{ for } C \in \mathrm{Circ}(a). \tag{6}$$

Note that $\prec_d$ is the majorization order and $p_1 \prec_c p_2$ is equivalent to $p_1 = c \star p_2$ for $c \in \mathrm{M}(a)$, where $\star$ denotes the circular convolution on $\mathbb{F}_a$.

Lemma 1 (Orders hierarchy)**.**

$$p_1 \prec_c p_2 \quad \Rightarrow \quad p_1 \prec_d p_2 \quad \Rightarrow \quad p_1 \prec_h p_2. \tag{7}$$

One can easily find examples showing that there is no reverse implications in Lemma 1. In this paper, we are interested in the $\prec_c$ order, and previous Lemma gives a first idea on how this order compares to the majorization order. Also note that the set of measures which are worst than a given $p \in \mathrm{M}(a)$ with respect to $\prec_c$ is given by the convex hull of the orbit of p through cycles, whereas it is given by the convex hull of the orbit of p through permutations when considering $\prec_d$. Note that if $p \in \widehat{\mathrm{M}}(a)$, these two sets are the same.

Definition 4. For $p \in \mathrm{M}(a)$, we define the Fourier transform of p by

$$\mathcal{F}(p)(\omega) = \sum_{k=0}^{a-1} p(k) e^{-2\pi i k\omega/a}, \quad \omega \in \mathbb{F}_a \tag{8}$$

and the inverse Fourier transform of $h : \mathbb{F}_a \to \mathbb{C}$ by

$$\mathcal{F}^{-1}(h)(k) = \frac{1}{a} \sum_{w=0}^{a-1} h(w) e^{2\pi i k\omega/a}, \quad k \in \mathbb{F}_a. \tag{9}$$

Remark 1.
1. $\mathcal{F}(p \star q) = \mathcal{F}(p)\mathcal{F}(q)$ for any $p, q \in \mathrm{M}(a)$.
2. If $p \in \widehat{\mathrm{M}}_k(a)$ with $p(k) = 1 - P$, we have that $\hat{p}$ is given by $\hat{p}(0) = 1$ and

$$\hat{p}(\omega) = (1 - \frac{aP}{a-1}) e^{-2\pi i k\omega/a}, \quad \omega \neq 0. \tag{10}$$

3. From previous remark, note that $(\widehat{\mathrm{M}}(a), \star)$ is a semi-group.

Definition 5. For $p \in \mathrm{M}(a)$, let $\mathrm{DOM}_c(p)$ be the set of probability measures which dominate p with respect to $\prec_c$, i.e., $\mathrm{DOM}_c(p) = \{q \in \mathrm{M}(a) : p \prec_c q\}$.

Remark 2. Note that it is easier to describe the set of measures that are dominated by a fixed measure p (cf. previous comment) than the reverse. However, we can write $\mathrm{DOM}_c(p) = \{q \in \mathrm{M}(a) : \mathcal{F}^{-1}(\mathcal{F}(p)/\mathcal{F}(q)) \geq 0\}$, and we can use the FFT algorithm to compute $\mathrm{DOM}_c(p)$ efficiently.

Lemma 2. *For any $a \geq 1$,*

$$p_1, p_2 \in \widehat{\mathrm{M}}(a), \quad p_1 \prec_h p_2 \quad \Rightarrow \quad p_1 \prec_c p_2. \tag{11}$$

Proof: Assume that $p_1 \in \widehat{\mathrm{M}}_k(a)$ and $p_2 \in \widehat{\mathrm{M}}_l(a)$ for $k, l \in \mathbb{F}_a$, and $p_1 \prec_h p_2$. Then, denoting $1 - P_1 = p_1(k)$ and $1 - P_2 = p_2(l)$,

$$\hat{p_1}(\omega)/\hat{p_2}(\omega) = \frac{1 - \frac{aP_1}{a-1}}{1 - \frac{aP_2}{a-1}} e^{-2\pi i (k \ominus_a l)/a}. \tag{12}$$

Hence, if $(1 - \frac{aP_1}{a-1})/(1 - \frac{aP_2}{a-1}) \in \mathrm{Im}(f)$, where $f : P \in [0,1] \mapsto (1 - aP/(a-1))$, we have that (12) is the Fourier transform of an element in $\widehat{\mathrm{M}}(a)$. This is easily verified since $\mathrm{Im}(f) = [0,1]$ and since by assumption $1 - P_1 \leq 1 - P_2$. ■

Since $\widehat{\mathrm{M}}(2) = \mathrm{M}(2)$, we have the following corollary.

Corollary 1.

$$p_1, p_2 \in \mathrm{M}(2), \quad p_1 \prec_h p_2 \quad \Rightarrow \quad p_1 \prec_c p_2. \tag{13}$$

IV. Universal source polarization

Definition 6 (Polar sensing sets)**.** Let $\varepsilon \in (0,1)$, $\ell \geq 1$ and $p \in \mathrm{M}(a)$, define $n := 2^\ell$,

$$\mathcal{S}_{\varepsilon,n}(p) := \{i \in [n] : H(U_i|U^{i-1}) \geq 1 - \varepsilon\}$$

where $U^n = G_n X^n$, $G_n = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]^{\otimes \ell}$, $X^n \overset{\text{iid}}{\sim} p$. We use the notation

$$\mathcal{S}(p_1) \subseteq \mathcal{S}(p_2) \text{ if } \mathcal{S}_{\varepsilon,n}(p_1) \subseteq \mathcal{S}_{\varepsilon,n}(p_2) \ \forall \varepsilon \in (0,1), n = 2^\ell.$$

Lemma 3. *For any $a \geq 1$,*

$$p_1 \prec_c p_2 \quad \Rightarrow \quad \mathcal{S}(p_2) \subseteq \mathcal{S}(p_1). \tag{14}$$

Proof: By assumption, there exists $c \in \mathrm{M}(a)$ such that $p_1 = p_2 \star c$. Let $X^n \overset{\text{iid}}{\sim} p_2$, $Z^n \overset{\text{iid}}{\sim} c$ independent of X^n and $\widetilde{X}^n = X^n \oplus Z^n \overset{\text{iid}}{\sim} p_1$. Define $U^n = G_n X^n$, $\widetilde{U}^n = G_n \widetilde{X}^n$ and $W^n = G_n Z^n$, hence $\widetilde{U}^n = U^n \oplus W^n$. We have

$$\begin{aligned} H(\widetilde{U}_i|\widetilde{U}^{i-1}) &\geq H(\widetilde{U}_i|\widetilde{U}^{i-1}, W^n) && (15) \\ &= H(U_i|U^{i-1}, W^n) && (16) \\ &= H(U_i|U^{i-1}) && (17) \end{aligned}$$

where the last equality follows from the fact that U^n is independent of W^n since X^n is independent of Z^n. ∎

Hence, form Lemma 3, we have that a polar code designed to compress a source for the distribution p_1, can equally well compress any source p_2 such that $p_1 \prec_c p_2$. (It will consume more rate than required for compressing a source under p_2 only, but it will allow lossless compression for both).

Assume that an i.i.d. source is generated from a distribution which is unknown, but which is known to belong to the set $\{p \in m(a) : h(p) \leq R\}$. We know (using method of types for example) that there exists a universal source code of rate R that can compress every i.i.d. sources in this set. Can we construct such a code using polar codes? It is ambitious to ask for such a "broad universality" with polar codes, since these are structured codes with complexity attributes, in contrast to the codes derived with the method of types. We may have to give up some extra rate to achieve this goal, or we may universally compress *only* certain subsets of source distributions. We now investigate these points.

A. Binary sources

For binary source, polar codes are universal in the following sense.

Proposition 1. *Any source which is known to have entropy at most R can be compressed universally with polar codes by storing the information bits on $\mathcal{S}(p^*)$, where p^* is one of the two distributions with entropy R. Moreover, if it is known on which symbol the source distribution puts more mass, the source can also be losslessly reconstructed.*

Proposition 2.

B. a-ary sources

Definition 7. For $D \subset \mathrm{M}(a)$, let

$$p_c(D) := \arg \min_{p \in \mathrm{M}(a) : p \prec_c D} H(p), \tag{18}$$

$$\hat{p}_c(D) := \arg \min_{p \in \widehat{\mathrm{M}}(a) : p \prec_c D} H(p). \tag{19}$$

In any of the above minimization, if the minimizer is not unique, pick one arbitrarily.

If the source distribution is known to belong to a set $D \subset \mathrm{M}(a)$, one way to construct a universal source code is to design a polar code for $p_c(D)$. Then, from Lemma 3 and Corollary 1, this polar code allows to compress losslessly any source in D. Of course, this may consume more rate than needed with an optimal source code, in other words, if we define $H_{\max}(D) := \max_{p \in D} H(p)$, we have in general

$$H(p_c(D)) \geq H_{\max}(D). \tag{20}$$

There are examples where equality holds in the above, in which case a polar code designed for $p_c(D)$ requires the minimal rate to compress losslessly the sources in D.

Lemma 4. *Let $D \subset \mathrm{M}(a)$ be such that $\arg\max_{p \in D} H(p)$ is unique (denoted $p_h(D)$) and satisfies $p_h(D) \prec_c D$. Then a source polar code designed for $p_h(D)$ can compress losslessly any source in D at the lowest achievable rate $\max_{p \in D} H(p)$.*

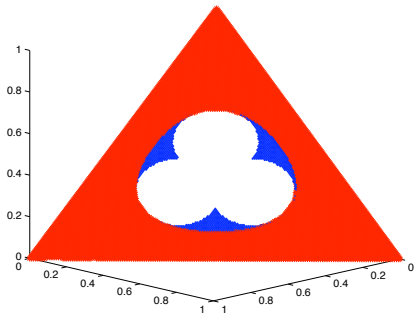

Fig. 1. Plots of $\mathrm{DOM}_h([0.2, 0.2, 0.6])$ included in $\mathrm{DOM}_c([0.2, 0.4, 0.4])$.

(Note that $p_h(D) \prec_c D$ implies that $p_c(D) = p_h(D)$.) The set $\mathrm{DOM}_c(p)$, plotted in Figure 1, satisfies (by definition) the hypothesis of Lemma 4 for any p. Comparing Figure 1 with the plot of $\mathrm{DOM}_h(p) := \{q \in m(a) : q \prec_h p\}$ also shows that there are sets for which (20) holds with a strict inequality.

An easy to construct such an example is to pick $D = p_1 \cup p_2$ where $p_1 \prec_h p_2$ but $p_1 \not\prec_c p_2$. Note that with such an example, we are not concluding that there exist sets D for which a universal source polar code does not exist, since with the ordering $\prec_c$, we may not consider the most general conditions to guarantee that $\mathcal{S}(p_2) \subseteq \mathcal{S}(p_1)$ (in other words, this inclusion may still hold if $p_1 \not\prec_c p_2$). However, one can attempt to show that there exist sets D for which a universal source polar code does not exist; this is investigated in Section IV-C. One can also check how much rate is lost by designing a polar code for a distribution that is dominated in terms of $\prec_c$ as opposed to $\prec_h$, for example the gap between the rate needed to compress $B_R := \{p \in \mathrm{M}(a) : H(p) \leq R\}$ using $p_c(B_R)$ and the minimal rate R needed with a random code, i.e., $H(p_c(B_R)) - R$. This gap can be computed using Remark 2 (e.g. it is 0.095 for R=0.865 and $a = 3$, case of Figure 1).

C. Non-universality of a-ary source polar codes and sensing via duality

In this section, we consider two source distributions p and q on $\mathbb{F}_a$, and we are interested in finding the rates at which one can compress these two sources with polar codes, allowing the reconstruction step to have knowledge on the source distribution. We denote by $C_{\text{pol}}(p, q)$ the infimum of these rates, and we provide different bounds on this quantity. This is a first step towards building polar codes which do not require knowledge of the source distribution at both compression and reconstruction steps, and yet achieve certain rates. Clearly

$$C(p, q) := H(p) \vee H(q) \leq C_{\text{pol}}(p, q).$$

From previous section, we have the upper bound $C_{\text{pol}}(p, q) \leq H(p_c(p, q))$, where $p_c(p, q)$ is as defined in (18) for the set $D = \{p, q\}$. In our definition, $C_{\text{pol}}(p, q)$ is given by the limit inferior of

$$\frac{1}{n} |\mathcal{S}_\varepsilon(p) \cup \mathcal{S}_\varepsilon(q)|.$$

While this is hard to compute, we can use the duality with channel coding as follows. Let $n = 2^\ell$, $U^n = G_n X^n$ where X^n is i.i.d. under p, and $V^n = G_n Y^n$ where Y^n is i.i.d. under q. Let us also denote by P (resp. Q) the additive noise channel whose noise distribution is p (resp. q). We then have from Section II

$$H(U_i|U^{i-1}) = 1 - I(P_i), \quad H(V_i|V^{i-1}) = 1 - I(Q_i) \quad (21)$$

where P_i (resp. Q_i) are the channels corresponding to P^σ for $\sigma \in \{-,+\}^\ell$, as defined in [1] with the tree construction. Moreover, if we define for $\varepsilon \in (0,1)$ $\mathcal{G}_\varepsilon(P) = \{i \in \{1,\dots,n\} : I(P_i) \geq \varepsilon\}$, we have

$$\mathcal{S}_\varepsilon(p) \cup \mathcal{S}_\varepsilon(q) = \left(\mathcal{G}_\varepsilon(P) \cap \mathcal{G}_\varepsilon(Q)\right)^c. \quad (22)$$

This shows that the compound capacity for source or channel coding are related and we can use the result of Section II and Theorem 5 in [7] to get the following bounds.

Lemma 5.

$$C_{pol}(p,q) \leq \frac{1}{2^\ell} \sum_{\sigma\in\{-,+\}^\ell} I(Z(P^\sigma)) \vee I(Z(Q^\sigma)) \quad (23)$$

$$C_{pol}(p,q) \geq \frac{1}{2^\ell} \sum_{\sigma\in\{-,+\}^\ell} H(p^\sigma) \vee H(q^\sigma) \quad (24)$$

where P (resp. Q) is the additive noise channel with noise distribution p (resp. q). Moreover each bound is monotonically approaching $C_{pol}(p,q)$.

Note that the upper bound is straightforward, and the notation $H(p^\sigma)$ refers to $H(U_i|U^{i-1})$ for the index i corresponding to σ. It is interesting to note that if BECs can be used to compute previous bounds, we cannot use the counter-example of [7] to show that polar codes do not achieve compound capacity in source coding, since BECs do not correspond to a valid source distribution via the duality of Section II. However, we can use the duality and BECs to construct sensing sets which are included in $\mathcal{S}_\varepsilon(p) \cup \mathcal{S}_\varepsilon(q)$, in a different manner than done in previous section. Let us give an example with $\ell = 1$. For two source distributions p and q, consider finding the BECs with parameter $Z(P)$ and $Z(Q)$ (P and Q as defined above). Then, as in [7], the good indices for P and Q satisfy

$$\mathcal{G}(P) \cap \mathcal{G}(Q) \supset \mathcal{G}(\text{BEC}(Z(P))) \cap \mathcal{G}(\text{BEC}(Z(Q))) \quad (25)$$

$$\equiv \mathcal{G}(\text{BEC}(Z(P) \vee Z(Q))) \quad (26)$$

and from (22), $\mathcal{G}(\text{BEC}(Z(P) \vee Z(Q)))$ gives a sensing set to compress losslessly p and q. This provides an interesting and different approach to constructing universal polar codes, although it may not be practical. In a work in progress, we propose the use of spike measures $\widehat{\text{M}}(a)$ to replace the "worst BECs" directly with "worst source distributions". The common feature between the spike measures and BECs is that they are both families that have a nested structures for the sensing/good index sets and that span the whole range of entropy/mutual information between 0 and 1. Also note that as opposed to the channel polarization case, degradedness in source polarization is less restrictive, since there are less degrees of freedom for source distributions than channels.

Now, to show that polar codes do not achieve the compound capacity in source coding, we can still use the lower bound of Lemma 5, but we need to pick two source distributions on ternary source alphabets.

Proposition 3. *Polar codes do not achieve the compound capacity for source coding when the source alphabet has strictly more than 2 elements.*

Counter-example: Let $p = [0.08, 0.36, 0.56]$, $q = [0.11, 0.62, 0.27]$, such that $H(p) = 0.8143$, $H(q) = 0.8126$ and $C = H(p) \vee H(q) = 0.8143$. The LHS of Lemma 5 for $\ell = 1$ evaluates at 0.8174 which is strictly larger than C.

V. Sparse signal recovery

In compressed sensing (CS), a sparse signal of high dimensionality can be recovered from a small number of random measurements with a convex optimization algorithm [5], [6]. More precisely, if $x \in \mathbb{R}^n$ is k-sparse (has at most k non-zero entries), and if

$$m = O(k \log n/k)$$

measurements are made of x using random projections, i.e., $y = \phi x$ where ϕ is an $m \times n$ random matrix with i.i.d. standard Gaussian entries, then x can be reconstructed within small ℓ_2 distance by searching for the sparse vector s minimizing $\|y - \phi s\|_{\ell_1}$.

In this section, we are interested in sensing a sparse signal using polar codes. As mentioned in the introduction, although similar in sprits, the source compression problem of previous section and the compressed sensing problem have a few distinctions: first, the source is a random process whereas the CS signal is deterministic; then the source is valued in $\mathbb{F}_q$ as opposed to $\mathbb{R}$ for CS. The second point can be addressed with quantization and is not discussed here. Moreover, many applications deal with signals which are valued in finite sets to start with. We hence focus here on the first point.

A possible way of defining k-sparse random sources, is to ask that the source distribution leads to an expected number of k non-zero values, or at most k, since it would be undesirable to have a sensing algorithm that does not work if the source is more sparse than expected. However, there are many distributions which would satisfy this assumption. Should a sensing algorithm succeed for all of them? Specifically, let a be a prime number and let $\mathbb{F}_a = \{0,\dots,a-1\}$. Let $X^n = (X_1,\dots,X_n)$ be i.i.d. samples from a distribution μ, with $\mu(i) = 1 - p$. Then, the number $K(X^n)$ of components of X^n which are not equal to i is in expectation

$$\mathbb{E}K(X^n) = np. \quad (27)$$

Let

$$\text{Spa}(\varepsilon) := \{\mu \in \text{M}(a) : \max_{i\in\mathbb{F}_a} \mu(i) \geq 1 - \varepsilon\},$$

and consider samples $X^n = (X_1,\dots,X_n)$ that are i.i.d. from a distribution in $\text{Spa}(\varepsilon)$. From previous remark, the number

of components in X^n that are not equal to a 'special value' is bounded by $n\varepsilon$. We could have considered $i = 0$ in the definition of $\mathrm{Spa}(\varepsilon)$, in which case we are considering sources that are sparse by having a bounded expected number of non-zero components[4]. However, the result we will derive does not depend on which component is frequent, and we do not restrict ourself to $i = 0$. This property may indeed be useful if the considered digital signal is constructed from a quantized signal that has been mapped to $\mathbb{F}_a$. On the other hand, $\mathrm{Spa}(\varepsilon)$ contains sources which can be supported on any subset of $\mathbb{F}_a$ (e.g., a may be large but this set still contains sparse binary sources), and it may be reasonable to assume that there is no such variation in the probability mass assigned to the non-special values. Hence, we also define

$$\widehat{\mathrm{Spa}}(\varepsilon) := \{\mu \in \widehat{\mathrm{M}}(a) : \max_{i \in \mathbb{F}_a} \mu(i) \geq 1 - \varepsilon\}, \qquad (28)$$

$$\widetilde{\mathrm{Spa}}(\varepsilon) := \{\mu \in \mathrm{M}(a) : s_{i,1-\varepsilon} \prec_c \mu, i \in \mathbb{F}_a\}, \qquad (29)$$

where $s_{i,1-\varepsilon}$ is the distribution of $\widehat{\mathrm{M}}_i(a)$ which has mass $1-\varepsilon$ at i and equally distributed elsewhere. Note that $\widetilde{\mathrm{Spa}}(\varepsilon) = \mathrm{DOM}_c(s_{0,1-\varepsilon})$ and $\widehat{\mathrm{Spa}}(\varepsilon) \subset \widetilde{\mathrm{Spa}}(\varepsilon) \subset \mathrm{Spa}(\varepsilon)$. The set $\mathrm{Spa}(\varepsilon)$ contains the corner of the simplex cut "straight", whereas $\widetilde{\mathrm{Spa}}(\varepsilon)$ contains the corner but cut in a diamond shape; hence $\widetilde{\mathrm{Spa}}(\varepsilon)$ does not contain some of the most "flat" distributions of $\mathrm{Spa}(\varepsilon)$ (such as the distribution assigning $1-\varepsilon$ at 0 and ε at 1).

Remark 3. Note that by symmetry, $p_c(\mathrm{Spa}(\varepsilon)) = \hat{p}_c(\mathrm{Spa}(\varepsilon))$ and hence, we do not lose optimality here by searching for the worst distribution in $\widehat{\mathrm{M}}(a)$.

Proposition 4. *Let X^n, with $n = 2^\ell$, be an n-sample drawn i.i.d. from a distribution in $D \subseteq \mathrm{Spa}(\varepsilon)$ and let ϕ be the $m \times n$ deterministic sensing matrix defined for D (cf. Definition 8). We have*

$$m = C_D k \log_a \frac{(a-1)n}{k} + o_\varepsilon(1), \quad k = n\varepsilon,$$

and with overwhelming probability, X^n can be exactly reconstructed from ϕX^n, using the polar decoding algorithm.

Remark 4.
1. The complexity of the polar decoding algorithm is $O(n \log n)$.
2. The multiplication ϕX^n is carried out over $\mathbb{F}_a$ and the logarithm's basis is a.
3. C_D is the cost to pay for having universality over D in the source distribution and it is a function of D, ε and a. If $D = \widehat{\mathrm{Spa}}(\varepsilon)$ or $\widetilde{\mathrm{Spa}}(\varepsilon)$, we have $C_D = 1$. If D is the entire set $\mathrm{Spa}(\varepsilon)$, C_D seems to increase very slowly with $1/\varepsilon$, as plotted in Figure 2.

Definition 8. Given a set D of probability measures on $\mathbb{F}_a$, we construct a sensing matrix for D and for the dimension $n = 2^\ell$ as follows:

(i) Find $p_c(D)$ as defined in (19)

[4]These may be the components of a real valued signal in an appropriate basis

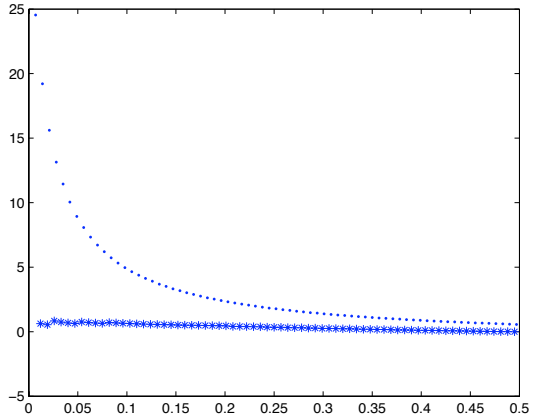

Fig. 2. Plots of C_D as a function of ε when $D = \mathrm{Spa}(\varepsilon)$, $a = 3$ (star-curve) and $a = 53$ (dot-curve).

(ii) Find $\mathcal{S} = \mathcal{S}_{\delta,n}(p_c(D))$ as in Definition 6 for $0 < \delta < 1$

(iii) Define $\phi = I_{\mathcal{S}} G_n$, where $G_n = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]^{\otimes \ell}$ and where $I_{\mathcal{S}}$ is the matrix whose columns indexed by $\mathcal{S}$ form the identity matrix and whose other columns are filled in with zeros. Note that ϕ is an $m \times n$ matrix, where $m = |\mathcal{S}|$.

Implementation of ϕ.
1. Step (i) can be easily computed, cf. Remarks 2 and 3.
2. Step (ii) requires an heavy computation: finding $\mathcal{S}$ with an analytic formula is a hard open problem in polar codes. However, it is mostly a mathematical challenge, since one can run simulations to determine $\mathcal{S}$ with a very good accuracy. Hence, step (ii) is also easily computed.
3. The construction of G_n is straightforward because of its Kronecker structure, and indeed, this structure is important to allow an efficient decoding algorithm running in $O(n \log n)$.

Polar decoding algorithm.
0. Initialize $\mathcal{M} = \mathcal{S}$.
1. For the smallest index i in $\mathcal{M}^c$, compute the likelihoods $\mathbb{P}\{(G_n X^n)_i = k | (G_n X^n)^{i-1}\}$, where $(G_n X^n)^{i-1}$ is known since we have sensed $G_n X^n$ on $\mathcal{S}$ to get ϕX^n. Decide for the most likely value of k and define $\mathcal{M} = \mathcal{M} \cup i$.
2. Iterate 1. until $i = n$.
3. Multiply $G_n X^n$ by $G_n = G_n^{-1}$ to get X^n.

As shown in [1], step 1. requires only $O(n \log n)$ computations. Since in each of these decisions, the entropy of the component to guess (given the past components) is close to 0, the true value is guessed correctly with high probability, and error propagation can be controlled. In a collaboration with V. Cevher and E. Telatar, the polar decoding algorithm is replaced with compressed sensing algorithms and numerical simulations suggest that the joint approach can succeed (work in progress).

REFERENCES

[1] E. Arıkan, *Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels,* IEEE Trans. Inform. Theory, vol. IT-55, pp. 3051–3073, July 2009.

[2] E. Arıkan and E. Telatar, *On the rate of channel polarization,* in Proc. 2009 IEEE Int. Symp. Inform. Theory, Seoul, pp. 1493–1495, 2009.

[3] E. Arıkan, *Source polarization,* in Proc. IEEE Int. Symp. Inform. Theory, Austin, 2010.
[4] E. Şaşoğlu, E. Telatar, E. Arıkan, *Polarization for arbitrary discrete memoryless channels*, August 2009, arXiv:0908.0302v1 [cs.IT].
[5] E. Candes, T. Tao , *Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies*, IEEE Trans. Inform. Theory, vol. 52, no. 12, pp. 5406-5425, December 2006.
[6] D. Donoho, *Compressed Sensing*, IEEE Trans. Inform. Theory, vol. 52, no. 4, pp. 1289-1306, April 2006.
[7] S. H. Hassani, S. B. Korada, R. Urbanke, *The Compound Capacity of Polar Codes*, arXiv:0907.3291v1 [cs.IT], July 2009.